What is the process for conducting truncated regression in Stata, and what does the annotated output reveal about the results?

Authored by

stats writer

June 30, 2024

The process of conducting truncated regression in Stata involves using the "truncreg" command to specify the dependent and independent variables, as well as the truncation point. This is followed by using the "regress" command to estimate the truncated regression model. The annotated output from this process provides information on the coefficient estimates, standard errors, and significance levels for the independent variables. It also includes a test for the assumption of normality and a diagnostic plot to assess the adequacy of the model. Additionally, the output reveals the value of the truncation point and the impact of the truncation on the regression results. Overall, the annotated output provides valuable insights into the results of the truncated regression analysis in Stata.

Truncated Regression | Stata Annotated Output

This page shows an example of truncated regression analysis with footnotes explaining the output. A truncated regression model predicts an outcome variable restricted to a truncated sample of its distribution. For example, if we wish to predict the age of licensed motorists from driving habits, our outcome variable is truncated at 16 (the legal driving age in the U.S.). While the population of ages extends below 16, our sample of the population does not. It is important to note the difference between truncated and censored data. In the case of censored data, there are limitations to the measurement scale that prevent us

from knowing the true value of the dependent variable despite having some
measurement of it. Consider the speedometer in a car. The speedometer may measure
speeds up to 120 miles per hour, but all speeds equal to or greater than 120 mph
will be read as 120 mph. Thus, if the speedometer measures the speed to be 120 mph,
the car could be traveling 120 mph or any greater speed-we have no way of
knowing. Censored data suggest limits on the measurement scale of the outcome
variable, while truncated data suggest limits on the outcome variable in the
sample of interest.

In this example, we will look at study of students in a special GATE (gifted
and talented education) program. We wish to model achievement (achiv) as
a function of gender, language skills and math skills (female,
langscore and mathscore in the dataset). A major concern is that

students require a minimum achievement score of 41 to enter the special program.
Thus, the sample is truncated at an achievement score of 40.

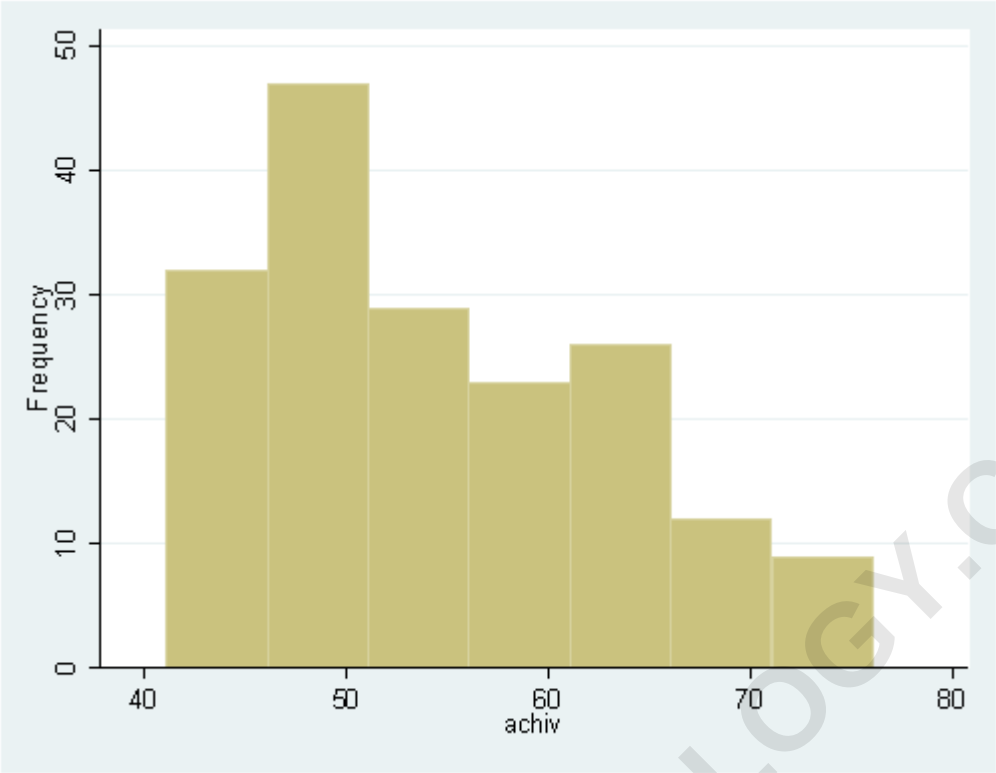
First, we can examine the data.

```
use  https://stats.idre.ucla.edu/stat/stata/dae/truncreg2, clear

summarize

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
id | 178 103.6236 57.08957 3 200
achiv | 178 54.23596 8.96323 41 76
female | 178 .5505618 .4988401 0 1
langscore | 178 5.401124 .8944896 3.1 6.7
mathscore | 178 5.302809 .9483515 3.1 7.4

histogram achiv, bin(7) freq
```



tabulate female

female | Freq. Percent Cum.

----------+----------------------------------

male | 80 44.94 44.94

female | 98 55.06 100.00

----------+----------------------------------

Total | 178 100.00

Now, we can generate a truncated regression model in Stata using the truncreg

command. We first list the outcome variable, then the predictors and the lower
and/or upper limit.

Our data are only left-truncated, so we will only indicate a lower limit,
ll(40).

truncreg achiv female langscore mathscore, ll(40)

(note: 0 obs. truncated)

Fitting full model:

Iteration 0: log likelihood = -580.98553
Iteration 1: log likelihood = -574.83026
Iteration 2: log likelihood = -574.53094
Iteration 3: log likelihood = -574.53056
Iteration 4: log likelihood = -574.53056

Truncated regression
Limit: lower = 40 Number of obs = 178
upper = +inf Wald chi2(3) = 89.85
Log likelihood = -574.53056 Prob > chi2 = 0.0000

----

achiv | Coef. Std. Err. z P>|z|

----------+----------

female | -2.290933 1.490333 -1.54 0.124 -5.211932 .6300663

langscore | 5.064698 1.037769 4.88 0.000 3.030709 7.098688

mathscore | 5.004054 .9555717 5.24 0.000 3.131168 6.87694

_cons | -.2940047 6.204858 -0.05 0.962 -12.4553 11.86729

----------+----------

/sigma | 7.739053 .5476443 14.13 0.000 6.66569 8.812416

Truncated Regression Output

(note: 0 obs. truncated)a

Fitting full modelb:

Iteration 0: log likelihood = -580.98553

Iteration 1: log likelihood = -574.83026

Iteration 2: log likelihood = -574.53094

Iteration 3: log likelihood = -574.53056

Iteration 4: log likelihood = -574.53056

Truncated regression

Limit: lowerc = 40 Number of obsf = 178

upperd = +inf Wald chi2(3)g = 89.85

Log likelihoode = -574.53056 Prob > chi2h = 0.0000

--

achivi| Coef.j Std. Err.k zl P>|z|m n

-------------+--

female | -2.290933 1.490333 -1.54 0.124 -5.211932 .6300663

langscore | 5.064698 1.037769 4.88 0.000 3.030709 7.098688

mathscore | 5.004054 .9555717 5.24 0.000 3.131168 6.87694

_cons | -.2940047 6.204858 -0.05 0.962 -12.4553 11.86729

-------------+--

/sigmao| 7.739053 .5476443 14.13 0.000 6.66569 8.812416

--

a.

(note: 0 obs. truncated)

- This indicates how many observations in the model had outcome variable values

below the lower limit or above the upper limit indicated in the function call.
In this example, it is the number of observations where achiv (leq) 40. The
minimum value of achiv listed in the data summary was 41, so there were
zero observations truncated.

b.
Fitting full model
- This is the iteration history of the truncated regression model. It lists the
log likelihoods at each iteration. Truncated regression uses maximum likelihood
estimation, which is an iterative procedure. The first iteration (called
Iteration 0) is the log likelihood of the "null" or "empty" model; that is, a
model with no predictors. At the next iteration (called Iteration 1), the
specified predictors are included in the model. In this example, the predictors
are female, langscore and mathscore. At each iteration, the log

**likelihood increases because the goal is to maximize the log likelihood. When
the difference between successive iterations is very small, the model is said to
have "converged" and the iterating stops. For more information on this process
for binary outcomes, see**

**Regression Models for Categorical and Limited Dependent Variables by J.
Scott Long (page 52-61).**

**c.
lower
- This indicates the lower limit truncation specified for the outcome variable. In this
example, the lower limit is 40.**

**d.
upper
- This indicates the upper limit specified for the outcome variable. In this
example, we did not specify an upper limit, so it is assumed to be infinity.**

e.

Log likelihood

- This is the log likelihood of the
fitted model. It is used in the Likelihood Ratio Chi-Square test of whether all
predictors' regression coefficients in the model are simultaneously zero.

f.

Number of obs

- This is the number of observations in the dataset where the outcome and
predictor variables all have non-missing values.

g.

Wald chi2(3)

-This is the Wald Chi-Square statistic. It is used to test the hypothesis that
at least one of the predictors' regression coefficient is not equal to zero. The
number in the parentheses indicates the degrees of freedom of the Chi-Square
distribution used to test the Wald Chi-Square statistic and is defined by the number of predictors

in the model (3).

h.

Prob > chi2

- This is the probability of getting a Wald test statistic as extreme as, or more so, than the observed statistic under the null hypothesis; the null hypothesis is that all of the regression coefficients across both models are simultaneously equal to zero. In other words, this is the probability of obtaining this chi-square statistic (89.85) or one more extreme if there is in fact no effect of the predictor variables. This p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. The small p-value from the test, <0.0001, would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero. The parameter of the chi-square distribution

used to test the null hypothesis is defined by the degrees of freedom in the
prior line, chi2(3).

i.
achiv
- This is the outcome variable predicted by the model.

j.
Coef. - These are the regression coefficients. They are interpreted in
the same manner as OLS regression coefficients: for a one unit increase in the predictor
variable, the expected value of the outcome variable changes
by the regression coefficient, given the other predictor variables in the model
are held constant.

female - The expected achievement score for a
female student is 2.290933 units lower than the expected achievement score for a
male student while holding all other variables in the model constant. In other
words, if two students, one female and one male, had

identical language and math
scores, the predicted achievement score of the male would be 2.290933 units
higher than the predicted achievement score of the female student.

langscore - This is the estimated regression
estimate for a one unit increase in langscore, given the other
variables are held constant in the model. If a student were to increase her
langscore by one point, her predicted achievement score would increase by
5.064698 units, while
holding the other variables in the model constant. Thus, the students with
higher language scores will have higher predicted achievement scores than students with lower
language scores, holding the other variables constant.

mathscore - This is the estimated regression
estimate for a one unit increase in mathscore, given the other
variables are held constant in the model. If a student

were to increase her
mathscore by one point, her predicted achievement score would increase by
5.004054 units, while
holding the other variables in the model constant. Thus, the students with
higher math scores will have higher predicted achievement scores than students with lower math scores, holding the other variables constant.

_cons - This is the regression estimate when
all variables in the model are evaluated at zero. For a male student (the
variable female evaluated at zero) with langscore and mathscore
of zero, the predicted achievement score is -0.2940047. Note that
evaluating langscore and mathscore at zero is out of the range of
plausible test scores.

k.
Std. Err.
- These are the standard errors of the individual

regression coefficients. They
are used in both the calculation of the z test statistic, superscript
l, and the
confidence interval of the regression coefficient, superscript n.

l.
z
- The test statistic z is the ratio of the Coef. to the Std. Err. of the respective predictor. The z value follows a standard normal distribution which is used to test against a two-sided alternative hypothesis that the Coef. is not equal to zero.

m.
P>|z|
- This is the probability the z test statistic (or a more extreme test statistic) would be observed under the null hypothesis
that a particular predictor's regression coefficient is zero, given that the
rest of the predictors are in the model. For a given alpha level, P>|z| determines whether or not the null

hypothesis
can be rejected. If P>|z|
is less than alpha, then the null hypothesis can be rejected and the parameter
estimate is considered statistically significant at that alpha level.

female - The z test
statistic for the predictor female is (-2.290933/1.490333) = -1.54 with an associated p-value of
0.124. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that the regression coefficient for
female has
not been
found to be statistically different from zero given langscore and
mathscore
are in the model.

langscore - The z test
statistic for the predictor langscore is (5.064698/1.037769) = 4.88 with an
associated p-value of <0.001. If we set our alpha level to

0.05, we would
reject the null hypothesis and conclude that the regression coefficient for
langscore has been
found to be statistically different from zero given female and
mathscore
are in the model.

mathscore - The z test
statistic for the predictor mathscore is (5.004054/0.9555717) = 5.24 with an
associated p-value of <0.001. If we set our alpha level to 0.05, we would
reject the null hypothesis and conclude that the regression coefficient for
mathscore has been
found to be statistically different from zero given female and
langscore
are in the model.

_cons - The z test
statistic for the intercept, _cons, is

(-0.2940047/6.204858) = -0.05 with an associated p-value of 0.962. If we set our alpha level at 0.05, we would fail to reject the null hypothesis and conclude that _cons has not been found to be statistically different from zero given female, langscore and mathscore are in the model and evaluated at zero.

n.

- This is the Confidence Interval (CI) for an individual coefficient given that the other predictors are in the model. For a given predictor with a level of 95% confidence, we'd say that we are 95% confident that the "true" coefficient lies between the lower and upper limit of the interval. It is calculated as the Coef. (zα/2)*(Std.Err.), where zα/2 is a critical value on the standard normal distribution.

The CI is equivalent to the z test statistic: if the CI includes zero, we'd fail to reject the null hypothesis that a particular

regression coefficient
is zero given the other predictors are in the model. An advantage of a CI is
that it is illustrative; it provides a range where the "true" parameter may
lie.

o.
/sigma
- This is the estimated standard error of the regression. In this example,
the
value, 7.739053, is comparable to the root mean squared error that would be
obtained in an OLS
regression. If we ran an OLS regression with the same outcome and
predictors, our RMSE would be 6.8549. This is indicative of how much the outcome
varies from the predicted value. /sigma approximates this quantity for
truncated regression.